位相的データ解析の現在 (統計的モデリングと予測理論のための統合的数理研究)

AUTHOR(S):

大林, 一平

CITATION:

大林, 一平. 位相的データ解析の現在 (統計的モデリングと予測理論のための統合的数理研究). 数理解析研究所講究録 2017, 2057: 34-50

ISSUE DATE:

2017-10

URL:

http://hdl.handle.net/2433/237190

数理解析研究所講究録
第2057巻 2017年 34-50

位相的データ解析の現在

大林一平 *
東北大学 原子分子材料科学高等研究機構
Ippei Obayashi
WPI-AIMR, Tohoku University

1 はじめに

本講究録では，位相的データ解析 (Topological Data Analysis, TDA) について概説する．TDA で重要なパーシステントホモロジーに注目し，その概要，数学的定義，現状の応用例，将来の課題等について述べる．

位相的データ解析とは，トポロジーの概念を用いてデータを解析する手法の総称である．データ解析では統計や機械学習のように確率論に基づいた手法，またフーリエ解析のように解析の概念をベースにした手法などが既存の手法としてよく使われてきた．ボロノイ図のような計算幾何学の手法は以前から利用されてきた[*1]．計算機によるホモロジーの計算が 1990 年代にある程度実用的になり，2000 年代にパーシステントホモロジーの概念が発展したことにより TDA というアイデアが注目を集めるようになった [2, 3, 4, 5, 6]．パーシステントホモロジーはターゲットとなる図形データに関し 1-パラメータ族を考えることでより多くの情報を得ることができるようになったもので，例えばこの 1-パラメータ族に空間スケールを割り当てることでデータのマルチスケールな幾何的情報を得ることができるようになった．すでにアモルファスの構造解析 [7, 8] やタンパク質 [9] の解析，ウイルスの遺伝的進化 [10] の解析，センサーネットワーク [11] の解析，などへの応用例などがある．

* 〒980-8577 宮城県仙台市青葉区片平 2-1-1，Email: ippei.obayashi.d8@tohoku.ac.jp

*1 ボロノイ図の数学的な概念自体は非常に古いものらしい．実用的な計算アルゴリズムである Fortune の走査線 (sweepline) アルゴリズムは 1986 年に発表された [1].

2　パーシステントホモロジーの概要

パーシステントホモロジーの応用で重要となるのが，パーシステント図である．典型的なパーシステント図の応用の場合，入力データは 2 次元もしくは 3 次元の有限個の点の集合 (ポイントクラウド) で，その出力がパーシステント図である．

パーシステント図の定義や計算方法はさておき，まず例題として [7] で述べられているアモルファスの構造解析の例について見ていこう．図 1 の (a) と (c) は分子動力学 (MD) シミュレーションで得た液体シリカとアモルファスシリカの原子配置を描画したものである．原子配置はどちらもランダムに見え，見た目で区別することは難しい．アモルファス構造と液体の原子レベルでの構造の違いを捉えるのは難しい，というのは良く知られた問題で，この 2 つの特徴付けは材料科学の重要な問題である．しかし原子配置から計算したパーシステント図 (図 1(b)(d)) ははっきりと異なり，簡単にこの 2 つを区別できるこれはこの 2 つの原子配置が (見た目ではわからないものの) 決定的に異なる幾何構造を持っている，ということを意味している．図 (b) と図 (d) を比べると，(d) のほうが特徴的な筋状のものが見える．これは後で詳しく議論するが原子配置の典型的構造が低次元的制約を持つということを意味している．

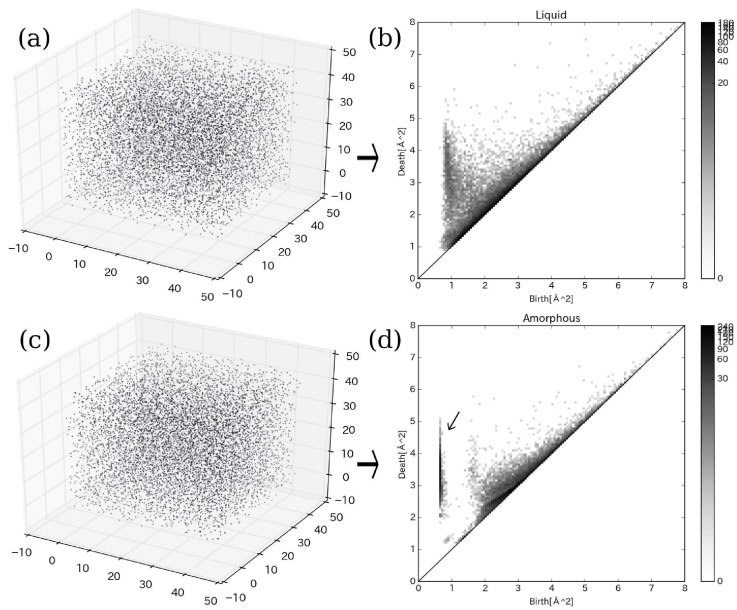

図 1　シリカの原子配置とそのパーシステント図 ([7] のデータを元に計算)

それでは，この図はどのようにして計算されるのであろうか．まず数学的基礎となるのが，アルファ複体とそのフィルトレーションである．入力となるポイントクラウドの位相の情報は明らかに「n 個の連結成分」だけである (図 2(a))．そのため，データの幾何的情報を見るために各点に半径 r の円を貼り付ける (図 2(b))．すると位相構造が新たに生成され，2 つの穴ができる (すなわち，$\dim H_1 = 2$ である)．こうすることによって図 (a) が「なんとなく」持っている構造を定式化することができる．この図 2(b) のような構造をアルファ複体と呼ぶ．

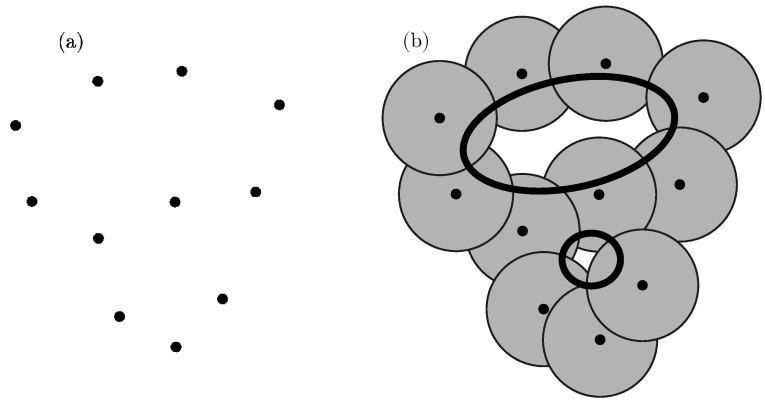

図 2　ポイントクラウドと円

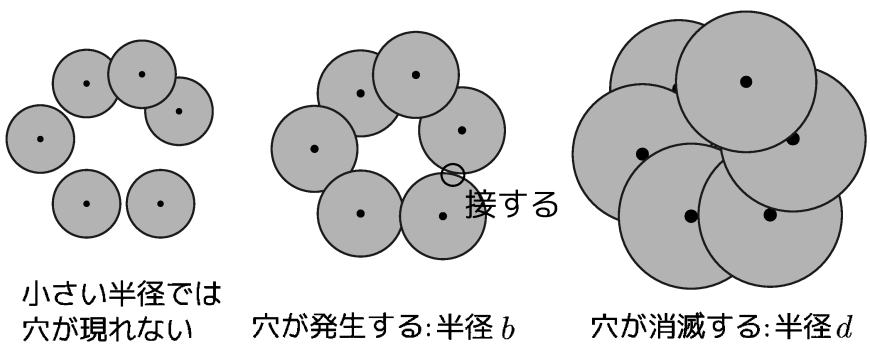

図 3　半径を変えたときの穴の生成と消滅

さて，ここで問題になるのが円の半径 r をいかに選ぶか，ということである．r によって穴が見えたり見えなかったりするのでこれは重要な問題である．ただ，この図 2 でいろんな半径を取ってみると，上側の穴は下側の穴より安定して存在する，ということがわかる．このような事実をうまく捉えるために，フィルトレーションの概念を用いる．半径 r

を 0 から大きくしていき，対象の図形の増大列を作る．すると，図 3 のように r が増大するにつれ穴が発生したり消滅したりする．パーシステントホモロジーの技法を用いると，各穴が発生する半径 b とそれが埋められて消滅する半径 d を対で計算することができる．この b を発生時刻 (birth time)，d を消滅時刻 (death time)，と呼び，対 (b,d) をパーシステンス対，生存対 (birth-death pair) と呼ぶ．最終的に得られるものは対の有限個の集合 $\{(b_i,d_i)\}_{i=1}^{s} \cup \{(b_i,\infty)\}_{i=s+1}^{s+t}$ [*2]であり，これを (x,y)-平面にプロットしたものがパーシステント図 (persistence diagram) である．ホモロジーは空間が $\mathbb{R}^N$ であれば 0 から $N-1$ の次数までは意味があるので，パーシステント図も普通 0 次から $N-1$ 次まで考える．0 次が連結成分，1 次が穴，2 次が空洞，の情報を持っている．

実は図 1 は 1 次のパーシステント図である．0 次や 2 次ではあまり特徴的な違いが見られないが，1 次では非常に特徴的な違いが見える．これはシリカガラスの構造がネットワークガラスと呼ばれるもので，共有結合によるネットワーク構造が重要であるために穴，すなわちリング状の構造を調べられる 1 次のパーシステント図に特徴が現れるのである[*3]．

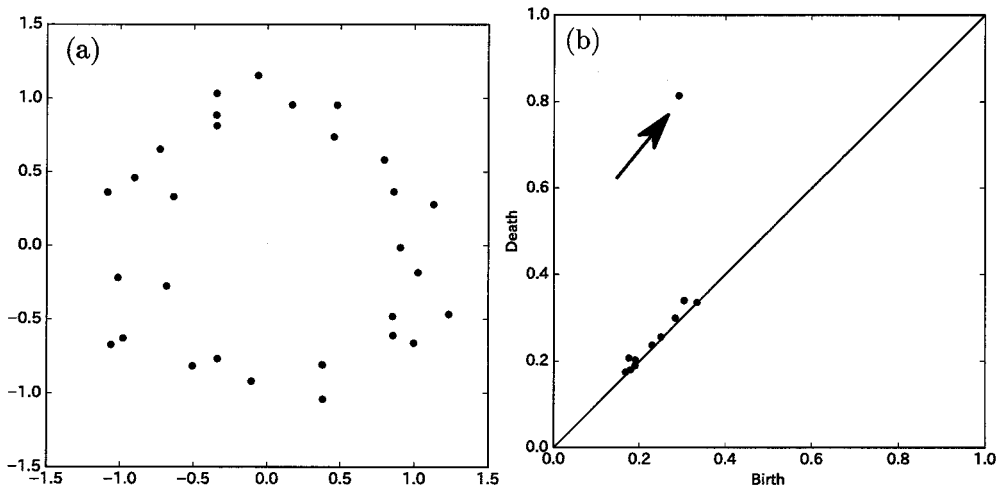

図 4　円形の点集合データ (a) とそのパーシステンス図 (b)

パーシステント図の意味をより深く考えるため，円形の点集合データ (図 4(a)) とその 1

[*2] 0 次のホモロジーを考えると，どんなに半径を大きくしても連結成分が一つ残る．これは消滅時刻が ∞ であるとみなす

[*3] ネットワーク型のアモルファスのリング構造を調べるというのは「リング統計」と呼ばれる手法が良く使われている．ある意味でパーシステントホモロジーによる解析はリング解析の発展版と見なせる．

次のパーシステント図 (図 4(b)) を調べよう．1 次のパーシステント図には対角線から離れた点が 1 つある．この点が図 4(a) の一番目立つリング構造を表現している．パーシステンス対の意味から考えるとその発生時刻 (x 軸) はリングをなす点の間の距離の 1/2 に対応し，その消滅時刻 (y 軸) はリングの半径に対応している．また，対角線のそばには数多くのパーシステンス対があるが，そのような対に対応するリング構造は発生してすぐに消滅するようなリングであり，すなわち点の位置をずらすとリングでなくなってしまうようなものが対応している．つまり対角線のそばのパーシステンス対はどちらかというとノイズ的なものであると言える．フィルトレーションを用いる，つまり半径パラメータ r を徐々に変化させることでこういったノイズ的なリングとしっかりした構造を持つリングを定量的に特定することが可能となる．また，対象となるデータに何らかの固有のスケール (例えば原子配置を考えるならば，原子間の距離の平均が重要なスケールになる) があるならば，それを目安にパーシステント図からノイズ的情報を除去することも可能である．

3 パーシステントホモロジーの数学的基礎

ここまでの議論は，数学的な厳密さや計算アルゴリズムの実現可能性などを避けてきた．この節では以下の内容について議論する．

- アルファ複体を単体複体として定義し，計算機上で実現しやすくする
- 単体複体の増大列からパーシステンス対を定義する
- パーシステンス対の計算アルゴリズム概要

3.1 アルファ複体

点集合の幾何構造の解析ツールとして，ボロノイ図が良く使われる．図 5 の 8 個の点に対し破線がその点集合のボロノイ図である．これは点集合 $P = \{u_i\}_{i=1}^N$ に対し，空間を「それぞれの点に最も近い領域」で分割したものである．そして「点集合の各点を頂点とする」「2 つのボロノイ領域が接するなら対応する点の間に辺を張る」「3 つのボロノイ領域が同時に接するなら対応する点の間に面を張る」「4 つ以上も同様」として構成された単体複体はドロネー複体と呼ばれる (図 5 の実線と灰色の部分，$\mathrm{Del}(P)$ と書く)．

半径パラメータ r に対応するアルファ複体 $\mathrm{Alp}(P,r)$ は $\mathrm{Del}(P)$ の部分複体として以下のように実現される [3].

- ドロネー複体の各頂点を頂点とする

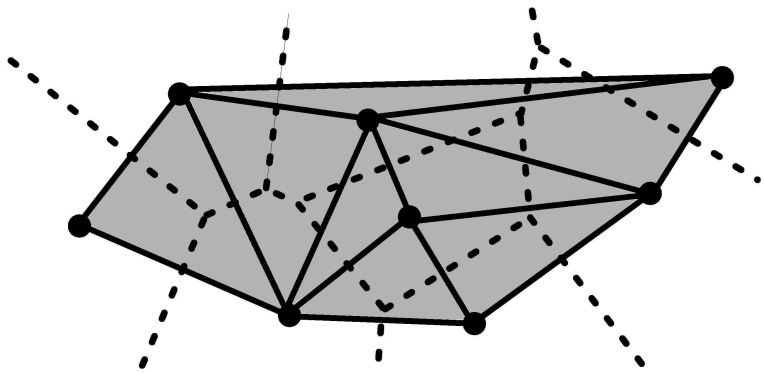

図 5　ボロノイ図とドロネー複体

- n 個のボロノイ領域が接するとき，領域に対応する n 頂点に半径 r の円盤を置いてその n 個の円盤が共通部分を持つならそこに $(n-1)$-単体を張る

図 6 では平面上の 3 点で r を変えた場合で，一方は面が張られ，もう一方は面が張られない例である．このようにして得られた単体複体は脈体定理 [4] より

$$|\mathrm{Alp}(P,r)| \simeq \cup_{u_i \in P} B_r(u_i)$$

が任意の $r > 0$ で成立する．ここで $\simeq$ はホモトピー同値を意味し，$|\cdot|$ は単体の幾何的実現を意味する．つまり半径 r の円を各点に置いたものとアルファ複体はホモトピー同値となり，ホモロジーの問題を考えるにあたっては $\mathrm{Alp}(P,r)$ を考えれば十分である．

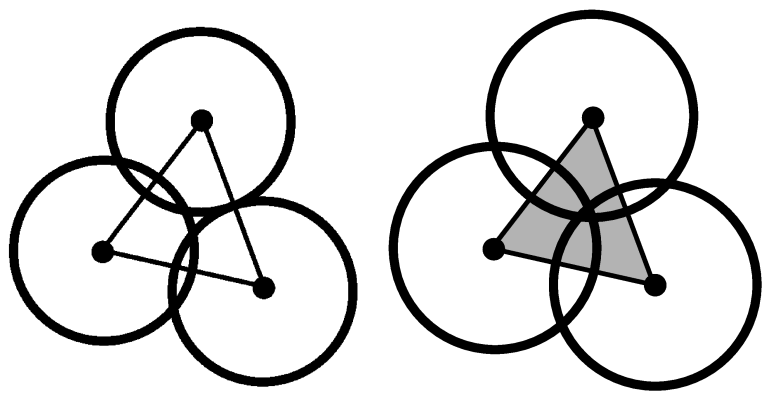

図 6　半径の大小によって面が張られない例と張られる例

実際に図 6 を良く見ると半径パラメータ r と 3 点の外接円半径との大小関係で中央の穴の存在/不存在が決まり，それとアルファ複体の定義が合致していることがわかる．ア

ルファ複体は有限単体複体であるので，計算機で表現するには好都合である．最終的に半径 r を 0 から増大させると，単体複体の増大列が得られる．これはアルファフィルトレーションと呼ばれる．また，ドロネー複体上の各単体 σ に対し，r を増大させたときにその単体が張られる瞬間の r を r_σ と書くことにする．アルファフィルトレーションは cgal (`http://cgal.org`) などで計算できる．

3.2 パーシステントホモロジー

実は，パーシステントホモロジーはアルファフィルトレーションに限らず一般の単体複体の増大列に対して定義できる．そこでまず単体 $\{\sigma_1, \ldots, \sigma_K\}$ からなる単体複体を考え，以下の条件が成立していると仮定しよう．

Condition 1. $\{\sigma_1, \ldots \sigma_k\}$ が任意の $1 \leq k \leq K$ において部分複体となる

このとき，$\{\sigma_1, \ldots \sigma_k\}$ の幾何的実現を X_k と置くと，$X_1 \hookrightarrow \cdots \hookrightarrow X_K$ という包含写像の列ができ，この列に対応する体 Q[*4]を係数とするホモロジー加群の列 $H_*(X_1) \to \cdots \to H_*(X_K)$ ができる．これをパーシステントホモロジーという．するとこの列はパーシステントホモロジーの構造定理 [6] により，既約区間 (interval decomposable)

$$\begin{aligned}\mathbb{I}[i,j) = 0 \to \ldots \to 0 \\ \to Q \xrightarrow{1} \ldots \xrightarrow{1} Q \to 0 \to \ldots \to 0 \\ (i \text{ 番目から } j-1 \text{ 番目まで非零})\end{aligned}$$

の直和で一意に分解できる．例えば

$$\begin{array}{ccccccc} H_\ell(X_1) & \to & H_\ell(X_2) & \to & H_\ell(X_3) & \to & \cdots \\ \| & & \| & & \| & & \\ 0 & \to & Q & \xrightarrow{\sim} & Q & \to & \cdots \\ \oplus & & \oplus & & \oplus & & \\ 0 & \to & Q & \to & 0 & \to & \cdots \\ \oplus & & \oplus & & \oplus & & \\ \vdots & & \vdots & & \vdots & & \end{array}$$

のような分解が可能となる．この分解において $0 \to Q$ はホモロジーの生成元の「発生」を，$Q \xrightarrow{1} Q$ はその「持続」を，$Q \to 0$ はその「消滅」を，それぞれ意味する．これが 2 節で説明した「穴」の生成と消滅の数学的に厳密な定義である．そして，各 $\mathbb{I}[i,j)$ に対して (i,j) という対を対応させたものがパーシステンス対である．

[*4] 議論の簡単さ，計算機での実装の容易さ，のため $\mathbb{Z}_2$ を使うことが多い．

アルファフィルトレーションは実数 r でパラメータ付けられているので，有限個での増大列の定義を直接適用することはできない．しかし，実質的に問題になるのは r が $\{r_\sigma \mid \sigma \in \mathrm{Del}(P)\}$ に含まれている場合のみであることがすぐにわかる．これは有限個であるのでドロネー複体の単体を r_σ によって昇順に並べ，同じ発生半径を持つ単体がある場合には適切に順序を決めると (この順で並べたものを $\{\sigma_1, \ldots, \sigma_K\}$ とする)，条件 1 を満たすようにできる．そして構造定理で得られる各既約区間 $\mathbb{I}[i,j)$ に対して，$(r_{\sigma_i}, r_{\sigma_j})$ を考えると，これがパーシステンス対となるのである．

この分解定理自体は体係数の多項式環が単項イデアル整域 (PID) であること，また，有限生成 PID 係数の加群は有限生成加群の構造定理より一意に分解可能であること，などから得られる帰結である[*5]．ここではパーシステントホモロジーの問題をどのようにして多項式環の問題に還元するか，という一般論は置いておいて，簡単な例題を代わりに用いることでこの分解について直感的に述べよう．ここでは，$f_1 : V_1 \to V_2, f_2 : V_2 \to V_3$ (V_1, V_2, V_3:有限次元ベクトル空間，f_1, f_2:線形写像) という写像があるとしよう．すると，以下のような分解が可能である．

$$
\begin{array}{ccccc}
V_1 & \xrightarrow{f_1} & V_2 & \xrightarrow{f_2} & V_3 \\
\| & & \| & & \| \\
\ker(f_1) & \to & 0 & \to & 0 \\
\oplus & & \oplus & & \oplus \\
\ker(f_2 \circ f_1)/\ker(f_1) & \xrightarrow{\sim} & \ker(f_2) \cap \mathrm{im}(f_1) & \to & 0 \\
\oplus & & \oplus & & \oplus \\
V_1/\ker(f_2 \circ f_1) & \xrightarrow{\sim} & \mathrm{im}(f_1)/(\mathrm{im}(f_1) \cap \ker(f_2)) & \xrightarrow{\sim} & \mathrm{im}(f_2 \circ f_1) \\
\oplus & & \oplus & & \oplus \\
0 & \to & \ker(f_2)/(\mathrm{im}(f_1) \cap \ker(f_2)) & \to & 0 \\
\oplus & & \oplus & & \oplus \\
0 & \to & V_2/(\mathrm{im}(f_1) + \ker(f_2)) & \xrightarrow{\sim} & \mathrm{im}(f_2)/\mathrm{im}(f_2 \circ f_1) \\
\oplus & & \oplus & & \oplus \\
0 & \to & 0 & \to & V_3/\mathrm{im}(f_2)
\end{array}
$$

この分解を見ればわかるように，分解定理は包含写像から導出されるホモロジー加群の列に限らず適用可能である．実際そういったアイデアによるパーシステントホモロジーの拡張 (例えばジグザグパーシステンス [12, 13]) が考案されている．

[*5] 体係数の多項式環が PID であることが本質的であるため，通常のようにホモロジーの係数は $\mathbb{Z}$ にはしない．

3.3 アルゴリズムの概要

この分解を計算するためには，上に挙げた分解に対応するようなうまいベクトル空間の基底を見つければよい．これは結局掃き出し法に帰結され，実際のアルゴリズムも掃き出し法に基いている．特に通常の単体複体の増大列に関するパーシステントホモロジーの分解は非常にシンプルかつ実用的なアルゴリズムが知られている．行列 B を基底 $\{\sigma_1, \cdots, \sigma_K\}$ に関する境界作用素 ∂ の行列表示である，としよう．すると条件 1 より B は上半三角行列となる．そこで「左の列のスカラ倍を右の列に加える」という操作のみで掃き出し法を進める．最終的に掃き出し法を進められなくなった段階での行列の各列が求める基底となり，求める分解が得られる．このアルゴリズムを [14] に述べられているように記述すると以下の通りである．

Algorithm 1 掃き出し法 (係数体が $\mathbb{Z}_2$ の場合)

for $j = 1, \ldots, n$ **do**
  while there exists $i < j$ with $\mathrm{low}(i) = \mathrm{low}(j)$ **do**
    add column i to column j

ここで $\mathrm{low}(i)$ は i 列で 0 でない要素の最大のインデックスであるとする (すべて 0 のときは $\mathrm{low}(i)$ は未定義とする)．たったこれだけのプログラムでパーシステンス対が計算できるのである．

このアルゴリズムは非常に簡単であるため，データ (行列) の表現法，並列化，といった様々な工夫が可能である．論文 [14] にはそういったアルゴリズムに関する話題や実際の計算ソフトウェアの性能評価などがなされている．

4 パーシステントホモロジーの応用例

この節では，パーシステントホモロジーの 2 つの応用例について詳しく紹介する．一つはウイルスの進化の解析 [10] で，もう一つはアモルファスの構造解析 [7, 8] である．これらの応用例によってパーシステントホモロジーの有用性を示す．

4.1 ウイルスの進化の解析

ここでは，[10] で述べられているウイルス進化のパーシステントホモロジーによる解析について紹介する．

ここで使われる道具は，ヴィートリス・リップス複体 (しばしばリップス複体と略される)[4] である．これまで議論してきたアルファ複体は，点に半径 r の円を貼り付けることで位相構造を構成している．しかし，ウイルスの遺伝子の研究をするにあたってはこの円という構造は意味を持たない．しかし遺伝子間の距離という概念は定義可能である．2 つの遺伝子の距離をゲノム配列の違いの大きさで計測する，というのは遺伝子の研究でよく使われるアイデアである．例えばハミング距離のようなもので距離を計測する．そこで，点の間の距離だけから定義できるリップス複体を以下のように定義する．

$$\mathrm{VR}(P,r) = \{\sigma = \{v_1,\ldots,v_k\} \subset P \mid d(v_i,v_j) \leq r \text{ for all } i,j\}$$

ここで $P = \{u_i\}$ は点の集合で，d は P 上の距離，$r > 0$ は距離パラメータである．r を変化させることでアルファフィルトレーションの場合と同様にフィルトレーションを構築できる．

遺伝子の進化に関しては 2 つの重要な要素がある．一つは「垂直伝播」で，親から子への遺伝情報の伝播である．ウィルスの場合は増殖過程でこれが行われる．もう一つは「水平伝播」で，ある個体が他の固体の遺伝子を取り込むことである．ウイルスの場合は遺伝子再集合 (reassortment) と呼ばれる現象があり，複数のウイルスが同じ宿主に感染した場合にそのウイルスの遺伝子が混ざり合う．ウイルスの場合には遺伝情報を守る壁 (細胞) がないため，このような混合現象が比較的頻繁に生じる，ということである．

ここで垂直伝播とランダムな遺伝子の変異しかない世界を考えてみよう．すると元々同じ遺伝子を持っていた個体が変異によって異なる遺伝情報を持つようになり，それが子孫へ伝えられさらに変異する，となる．このようにして遺伝子の多様性が増大し，遺伝子の距離による構造は木構造になる．遠い祖先で異なる遺伝子を持つようになった 2 つの個体は距離が大きく，近い祖先で異なる遺伝子を持つようになった 2 つの個体は距離が近くなる．これによって木構造が実現される．そして木構造の 1 次のホモロジー加群は自明となるため，この世界では遺伝子距離による 1 次のホモロジー加群の次元は 0 に近くなるはずである．しかしここで水平伝播という概念が加わると事情は変化する．水平伝播が生じることで現れる個体の遺伝子は，その元となった 2 つ (以上) の個体の遺伝子のどれともそれなりに近しいはずである．つまり水平伝播によって木構造の異なる枝がつながるという現象が生じる．これによって枝の間に橋ができることで 1 次のホモロジーが現れる．これをパーシステントホモロジーで解析するのである．

ここで紹介している論文ではシミュレーションでの遺伝子の変化の様子と実際のインフ

ルエンザウイルスの遺伝子データをパーシステントバーコード*6 を用いて比較することで，例えば遺伝子再集合が生じる頻度を推定したり，どのようなパターンが特徴的に頻繁に生じるかを議論したりしている．

4.2 アモルファスの構造解析

次に述べるのは [7, 8] でのアモルファスの構造解析である．2 節で少し紹介したが，この節でより詳しく解説する．

アモルファスは非晶質とも呼ばれ，結晶のような繰り返し構造 (長距離秩序，Long Range Order と呼ばれる) は存在しないが，最近接原子や隣の隣の原子までの距離といった短距離では構造を持つ (短距離秩序，Short Range Order と呼ばれる) ような固体である．代表的な物質としてシリカガラス (SiO_2) がある．一方液体も短距離秩序がある一方で長距離構造を持たず，アモルファスと液体の違いは何なのかははっきりとはわかっていない．現在のところ，アモルファス特有の構造として中距離秩序 (Midium Range Order) と呼べるものが存在するのではないかと考えられている．ここで紹介する論文では，この中距離秩序をパーシステント図で特徴付けている．

この論文の内容を紹介するために，まず optimal cycle[15, 16] について紹介する．ホモロジーはチェイン複体 $\{C_q\}_q$ 上の境界作用素 $\partial_q : C_q \to C_{q-1}$ によって $H_q = Z_q/B_q, Z_q = \ker\partial_q, B_q = \mathrm{im}\partial_{q+1}$ と定義される．H_q の基底ベクトルが図形上の穴や空洞などを表している．しかし，B_q で割っているため，この基底ベクトル自体は具体的な幾何的オブジェクトではない．例えば図 7 に対しては $\dim H_1 = 1$ であるが，$z_i + B_1 (i = 1, 2, 3)$ はすべてこの H_1 の基底である．しかしこのなかで z_3 がこの穴を一番よく表していると考えられる．この z_3 の特徴付けとして何が良さそうかを考えると，「ループの長さが最も短い」のが良いのではないかと考えられる．このような z_3 のことを optimal cycle と呼ぶ．基本的にこのような optimal cycle を探す問題は線形空間上のある種の最適化問題として定式化され，その計算アルゴリズムが良く調べられている．[15] は通常のホモロジー加群に関する optimal cycle についての論文であり，[16] によってパーシステントホモロジーに関する問題として拡張されている．この optimal cycle を用いるとパーシステント図の上の各パーシステンス対が元データのどの部分にあたるのかを知ることができる．optimal cycle はここで解説するアモルファスの解析に有効利用されている．

図 8 は [7] より引用した，液体シリカ/アモルファスシリカ/結晶シリカの原子配置より

*6 パーシステントホモロジーの可視化手法の一種．パーシステント図と同じ情報を別の形で可視化する．

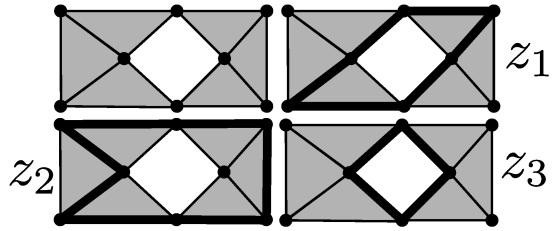

図 7　Optimal Cycle の問題

計算された 1 次のパーシステント図である．結晶の場合は，パーシステンス対はいくつかの小さい島状に分布し，液体の場合は，2 次元的に分布が広がっている．アモルファスの場合はどちらとも異なり，曲線的な分布をしている．パーシステント図を用いることでアモルファスを結晶，液体の双方から区別することができたのである．アモルファスシリカは共有結合による乱雑なネットワーク構造が主要な構造である (このようなガラスをネットワークガラスと呼ぶ) ので，1 次のパーシステント図が有用なのである．

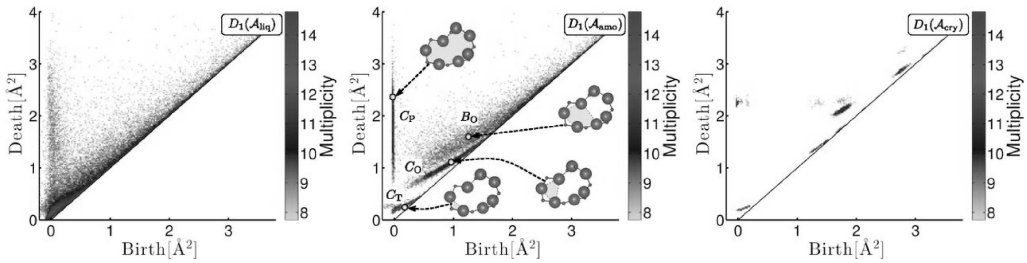

図 8　液体シリカ/アモルファスシリカ/結晶シリカの 1 次のパーシステント図 ([7], Fig 2 より引用，カラーの図表をグレイスケール化している)

さらにこの分布の意味するところを調べるために optimal cycle が活用される．図 8 の中央のアモルファスに対する図には，パーシステント対に対する optimal cycle を示している (大きい丸が酸素原子，小さい丸がケイ素原子を表している)．C_T は最近接の 3 原子がなす構造と対応しており，これは上で言う短距離秩序に対応している．C_O，B_O，C_P はより離れた原子からなる構造を表していることがわかる．これが中距離秩序に対応する構造であろうと考えられる．

曲線的な分布については，パーシステント図の安定性の概念と関連している．入力のポイントクラウドを連続的に動かしたとき，パーシステント図上の各パーシステンス対は連続的に変化することが知られている．図 8 の C_O は酸素原子 3 個からなる構造が典型的に原子配置上に現れていることを意味するが，さらにそれが曲線的分布をしているというこ

とは，この安定性の概念から典型的配置が低次元的な制約を受けているということを意味することがわかる．実際この3原子を取り出してきてそれがなす三角形の二辺の長さとその間の角度をプロットすると低次元的構造を持つことがこの論文では示されている．曲線的分布の幅はその低次元構造からのずれ具合を表している．

この論文ではパッキングガラスと呼ばれる金属原子によるガラスで典型的な構造も扱っている．半径のそろった原子を3次元空間につめこむと，典型的には体心立方格子 (BCC)，面心立方格子 (FCC)，六方最密充填 (HPC) といった結晶構造をなす．しかし大きさの異なる原子が3次元空間に詰め込まれることでランダムなばらつきが生じ，アモルファスとなる．この場合のアモルファスの特徴付けは1次のパーシステント図ではなく主に2次のパーシステント図でできることが示されている．球による空間充填構造が重要なので2次のパーシステント図が有効なのである．

このように，様々な種類のアモルファスがパーシステント図によって特徴付け可能である．ネットワークガラスとパッキングガラスの違いがパーシステント図の次数の違いに現れるというのも興味深い．

5　その他の問題など

ここでは，上の応用例では説明できなかったTDAの広がりについて簡単に解説する．

5.1　統計手法や機械学習との連携

この節では統計や機械学習の手法をいかにパーシステントホモロジーの技法と結び付けるかについての研究の現状について簡単に紹介する．

統計手法や機械学習は現在のデータ分析の主流の手法である．データから法則，パターンを見つけだし，新たなデータにそのパターンをあてはめることでその性質を推定したりすることができる．統計や機械学習では (例外もあるが) 通常入力としてベクトルを取る．元データをなんらかの方法でベクトル化してから学習させることが多い．このベクトル化の善し悪しが機械学習の性能を決めることも多い[*7]．

パーシステント図を統計や機械学習で活用するためには，パーシステント図をベクトルに変換するのが最も自然である．元々のデータに含まれる幾何的特徴がパーシステント図に効率的に抽出されて含まれていると考えられるため，パーシステント図を経由することでより良い処理ができるはずである．現在，ベクトル化には以下のような様々な手法が提

[*7] このベクトルは「特徴量」「記述子」「素性」などと呼ばれる．分野によってどの用語を使うかが異なる．

案されている．

- Persistence landscape[17]
- Persistence Scale Space Kernel (PSSK)[18]
- Persistence Weighted Gaussian Kernel (PWGK)[19]
- Persistence image[20]

今のところ決定版と呼べる手法はなく，様々な手法が試されている．基本的にはどれも適当な L^2 関数空間に埋め込む手法である．[18] では 3 次元形状データの分類問題への応用がなされ，[19] では上の応用例と同様のアモルファスの原子配置データを用いた液体からガラスへの転移点の検出などが応用としてなされている．

5.2 多次元パーシステンス

ここまでの解説では，パーシステントホモロジーは一列に並んだ単調増大列を考えてきた．しかし，二次元的な増大列を考えることができるならばより便利であろう．つまり，以下の図のような 2 軸の包含関係を考えるということである．

$$\begin{array}{ccccc} X_{1,1} & \hookrightarrow & X_{1,2} & \hookrightarrow & \cdots \\ \uparrow & & \uparrow & & \cdots \\ X_{2,1} & \hookrightarrow & X_{2,2} & \hookrightarrow & \cdots \\ \uparrow & & \uparrow & & \cdots \\ \vdots & & \vdots & & \end{array}$$

[21] で述べられている応用例としては，一つの軸はこれまで利用してきた，貼り付ける円を大きくする方向，もう一つはノイズ除去の強さによる方向，とするものである．パーシステントホモロジーは点がずれるようなノイズには強いものの，外れ値には弱いという性質がある．そこで外れ値の検出技法によってそのような点を除去することを考える．しかしここでノイズ除去の強さパラメータを決めるという問題が新たに生じる．そこでこのパラメータもフィルトレーションに含めてしまおう，というアイデアである．

このためには，以下のようなホモロジーの可換図式の中にある構造を一次元パーシステントホモロジーの構造定理のような方法で同定できればよいのである．

$$\begin{array}{ccccc} H_\ell(X_{1,1}) & \to & H_\ell(X_{1,2}) & \to & \cdots \\ \uparrow & & \uparrow & & \cdots \\ H_\ell(X_{2,1}) & \to & H_\ell(X_{2,2}) & \to & \cdots \\ \uparrow & & \uparrow & & \cdots \\ \vdots & & \vdots & & \end{array}$$

しかし，[21] で述べられている事実は，そのような構造を一般的に見つけだすのは二次元以上の場合では不可能である，ということである．しかしそうはいっても多次元のパーシステントホモロジーを考えることに需要はある．そこで考えられるアイデアとしては

- この可換図式の構造を部分的に説明できるような道具を探す
- 図式の形状を限定して考える
- よくあるデータ，典型的データでうまくいく道具立てを考える

[21] では一つ目のアプローチにより rank invariant という概念を提案している．RIVET[22] ではこのアイデアを使って多次元パーシステントホモロジーの見事な可視化ソフトウェアを実現している．

[23] では二つ目のアプローチで commutative ladder という名前で図式が小さな梯子型 (2×2, 2×3, 2×4) ならば構造を計算できる，ということを示している．この証明には表現論のツールなども使われている．

三つめのアプローチの成功例は寡聞して聞いたことがないが，データ処理の文脈から言えばこういったアイデアも何かあるのではないだろうか．例えば行列の対角化を考えると，一般的にはジョルダン標準形などを考えなければすべての場合にうまくいくとはいえないが，実際問題としては対象となる行列は対角化可能だろうと仮定して数値計算することが多い．また，対称行列に限定して対角化したい，といったときにはより効率的なアプローチもできるわけである．正直あまり良いアイデアがあるわけではないが，こういう方向性も有望ではないだろうか．

6 おわりに

この講究録で述べられてきたように，パーシステントホモロジーは様々な応用がある．さらにウイルスの遺伝子の解析でとアルファ複体ではなくリップス複体を用いたり，またアモルファスの解析で optimal cycle という概念が有効であったように，データからちょっとパーシステント図を計算すればおしまい，というものではなく，様々な工夫や新たな数学的概念が必要となる．多次元のパーシステントホモロジーのように数学的な困難を持った問題も多い．多次元のパーシステントホモロジーをやろうと思うと表現論が顔を出したり，ここでは説明しなかったが圏論が有効だったりもする．このようにパーシステントホモロジーと他の分野の数学との繋がりも見えてきた状況である．統計や機械学習との連携もまだ始まったばかりである．新たな応用に対しては新たな理論的道具が必要とされるはずである．パーシステントホモロジー，位相的データ解析はまだまだ理論的問題や応用の

問題が山積みであり，これらは今後も発展しつづけると筆者は期待している．

参考文献

[1] S. Fortune. A sweepline algorithm for Voronoi diagrams. Proceedings of the second annual symposium on Computational geometry. 313–322. (1986)

[2] G. Carlsson. Topology and Data. Bull. Amer. Math. Soc. **46** (2009), 255–308.

[3] H. Edelsbrunner and J. Harer. Computational Topology: An Introduction. AMS (2010).

[4] 平岡裕章, タンパク質構造とトポロジー -パーシステントホモロジー群入門-, シリーズ・現象を解明する数学 (三村 昌泰・竹内 康博・森田 善久 編集), 共立出版 (2013).

[5] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological Persistence and Simplification. Discrete Comput. Geom. **28**(4) (2002), 511–533.

[6] A. Zomorodian and G. Carlsson. Computing Persistent Homology. Discrete Comput. Geom. **33**(2) (2005), 249–274.

[7] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue, and Y. Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. PNAS 2016 **113** (26), 7035–7040

[8] T. Nakamura, Y. Hiraoka, A. Hirata, E. G. Escolar, and Y. Nishiura. Persistent homology and many-body atomic structure for medium-range order in the glass. Accepted in Nanotechnology.

[9] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda. Topological measurement of protein compressibility via persistent diagrams. Japan J. Indust. Appl. Math. **32** (2015), 1–17.

[10] J. M. Chan, G. Carlsson, and R. Rabadan. Topology of viral evolution. PNAS **110**(46) (2013), 18566–18571.

[11] V. de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. Algebraic and Geometric Topology **7** (2007), 339–358.

[12] G. Carlsson, V. de Silva, and D. Morozov. Zigzag Persistent Homology and Real-valued Functions. Proceedings of the Annual Symposium on Computational Geometry (2009) 247–256.

[13] G. Carlsson and V. de Silva. Zigzag Persistence. Foundations of Computational Mathematics **10**(4) (2010), 367–405.

[14] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington. A roadmap

for the computation of persistent homology. http://arxiv.org/abs/1506.08903.

[15] T. K. Dey, A. N. Hirani, and B. Krishnamoorthy. Optimal homologous cycles, total unimodularity and linear programming. SIAM Journal on Computing **40**(4) (2011), 1026–1044.

[16] E. G. Escolar, and Y. Hiraoka, Optimal Cycles for Persistent Homology Via Linear Programming, Optimization in the Real World: Toward Solving Real-World Optimization Problems, Springer Japan (2016), 79–96.

[17] P. Bubenik. Statistical topological data analysis using persistence landscapes. The Journal of Machine Learning Research archive **16**(1) (2015), 77–102.

[18] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A Stable Multi-Scale Kernel for Topological Machine Learning. IEEE Conference on Computer Vision and Pattern Recognition (2015) 4741–4748.

[19] K. Genki, K. Fukumizu, and Y. Hiraoka. Persistence weighted Gaussian kernel for topological data analysis. Proceedings of the 33rd International Conference on Machine Learning (ICML) (2016).

[20] H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, and L. Ziegelmeier. Persistence Images: A Stable Vector Representation of Persistent Homology. http://arxiv.org/abs/1507.06217.

[21] G. Carlsson and A. Zomorodian. Discrete and Computational Geometry **42** (2009) 71–93.

[22] M. Lesnick and M. Wright. RIVET. http://rivet.online/.

[23] E. G. Emerson and Y. Hiraoka. Persistence Modules on Commutative Ladders of Finite Type. Discrete and Computational Geometry **55** (2016) 100–157.